Open Data Lab Annual Report - 2018

The Open Data Lab Collaboration

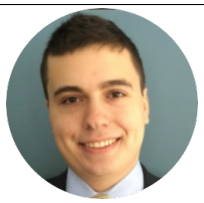
March 6, 2019

The Team

Pete Alonzi came to data science by way of the particle physics community. As a result he has great interest in making data open and usable to broad audiences. He sereves as a data scientist at the DSI and is the project manager for the Open Data Lab.

Phil Bourne "At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excuri sint occaecati cupiditate non provident,At vero eos et accusamus et dent, "

Tim Clark "Tim Clark, Ph.D., is a biomedical informatician and computer scientist with 28 years experience in academic, government and industry. He is an expert in biomedical knowledge representation, data fusion and open science applications. He is an Associate Professor appointed in the UVA School of Medicine & the Data Science Institute."



Max Levinson "Max Levinson is a Software Engineer and Cloud Developer within Public Health Services. When not building out REST apis for the NIH Data Commons Project, he can be found hacking on the Open Data Lab Project.
Max is passionate about microservices development, knowledge graphs, and anything python."

Daniel Mietchen "At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excuri sint occaecati cupiditate non provident,At vero eos et accusamus et dent, "

Lane Rasberry "At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excuri sint occaecati cupiditate non provident,At vero eos et accusamus et dent, "

Letter from the editor

hello world

Contents

Chapter 1

Overview

1.1 What is the Open Data Lab?

OPEN

We encourage all users to be as open as possible with every aspect of their work. That may be in opening up their data sets, publication, source code, or ...

DATA

We take an expansive definition of data. Everything from traditional data, to code, to workflows, to published material, and so on is considered data to us. We provide a place for all things digital data.

LAB

We provide a place where the power of contemporary computing can be brought to bear against data resources. Given the scale of data today this means colocating the data and computational resources.

Open Data Lab

The Open Data Lab is a resource to provide state of the art computing and data infrastructure to researchers, students, and sharers. It is guided by the principles of science and openness.

1.2 User Archetypes

There are many potential use cases for the Open Data Lab. In this section we describe the three cases that have been tested so far. They are: the Collaborator, someone who is working on a research project; the Student, someone who is using the Open Data Lab to learn about data science; and the Sharer, a person with data who wants to open it up to a broader audience.

The Collaborator

This archetypal person uses the Open Data Lab to conduct research. They access data and computational resources that are colocated. This colocation facilitates lower latency and increased performance. A wide range of services can be provided globally by AWS and locally through UVA HPC resources. Sample workflow:

1. Request a user account on the Open Data Lab
2. Once per collaboration:
   (a) Load data
   (b) Provision computational resources
3. Conduct research operations
4. Register resulting products in Dataverse

The Student

This archetype uses the Open Data Lab to facilitate learning. An example would be someone who participates in a workshop where and ODL notebook instance powered by AWS SageMaker provides the working environment. Sample workflow:

1. Request a user account on the Open Data Lab
2. Logon to AWS console to launch Jupyter
3. Use Jupyter during the workshop

The Sharer

This archetype is a user who owns data and wants to make it available. There are many mechanisms for sharing the data ranging from RESTful API of S3, to a SPARQL endpoint. Sample workflow:

1. Request a user account on the Open Data Lab
2. Load data into an S3 bucket
3. Configure one of the following
   (a) SPARQL endpoint
   (b) API Gateway to access S3
   (c) S3 permissions for a SageMaker notebook
   (d) ...

1.3 User Summary

group	projectID	# members	type
Bourne-Mura	bamc	4	MSDS Capstone
CBW	cbwc	4	MSDS Capstone
Wiki	wiki	9	MSDS Capstone
Mental Health	miip	6	SYS Capstone
Women Terror Recruitment	watr	2	Presidential Fellow
Healthy Markets	hmtt	5	DSI Research
Independent Study	pmis	1	DSI Research
Linked Open Data	nept	2	External Data
Spark	sprk	17	Education
GitHub	gith	9	Education
ORCI	orci	2	ODL Development
ML under	mlunder	7	Club
ML grad	mlgrad	3	Club
Rivanna	–	11	Local
Ivy	–	6	Local
ODL-education	–	26	Education Users
ODL-users	–	68	Unique Users

1.4 A phased approach

The first three phases of the Open Data Lab have been outlined. Phase 0 focused on pre investigation and decided on what technology to test in Phase 1, the closed beta. Phase 2 is an open Beta and will serve the community of Charlottesville and other associated research and educational efforts.

Phase	type	start	end
0	alpha	FEB 2018	JUN 2018
1	closed beta	JUL 2018	MAY 2019
2	open beta	TBD	–

Short Term Goals

1. Explore technology to solve open Open Data Lab questions
2. Serve Data Science Institute users
3. Understand user archetypes

Long Term Goals

Change the way science is done.

1.5 What's next for the Open Data Lab?

The year 2019 will be a big year for the Open Data Lab. The gathering of information and skill from the closed beta has gone well. While continuing the closed beta through the spring semester 2019 the Open Data Lab will be seeking to add staff to prepare for the open beta launch.

Chapter 2

Key Developments

This year saw great progress for the Open Data Lab. Early on there were weekly meetings to define goals and objectives. Those meetings lead to implementation of several areas and this chapter presents the key developments during 2018.

2.1 Phase 1 Closed β

The first decision made for the Open Data Lab was to implement a staged approach. Section 1.4 summarizes the whole scope. This section discusses the state of the closed beta (Phase I). The user base for this phase is predominantly the Data Science Institute. There were 42 participants ranging from students, to faculty, to staff. There were also 26 workshop attendees ranging from a diverse selection of UVA researchers to community members. The primary goal of this phase is to test different technology solutions to anticipated needs (see section 2.3. Those range from data storage to computation to discovery to pedagogy, and so on. Of particular note was the wild success of implementing Project Jupyter. The tools developed by this project served many roles and were excelent (details in section 2.3.1. This phase will run until several criteria are met. The first is the establishment of a new funding model that will cover the scope of the open beta test. The second is the acquisition of new staff. Currently the Open Data Lab has a bus factor of one and that is not acceptable for phase 2. There are other criteria to be developed the chief of which is to scope the open beta.

2.2 Establishment of User base

This year saw the birst of the Open Data Lab and growth to include 68 users. Those users take various form from capstone research programs at the graduate and undergraduate level to full fleged dissertation research. Some of the users involved are dealing with datasets that now reside within the Open Data Lab. Currently those datasets and under tight restriction as we explore proper security protocols. There is also a contingent of undergrad and graduate students that are part of data science clubs at UVA who gain access to resources through the Open Data Lab.

It is important to understand that the technology behind the system is not the driver of the system. The needs of the user are. Right now the closed beta format allows us to interview new users and tailor a program to them. Sometimes we get the resource wrong and adjustments have to be made. Regarding data storage the use of S3 storage from Amazon Web Services has served a broad selection of users well. Recent developments in AWS object storage technology enable users to use it as if it were block storage. As a result S3 has proved an effective solution both for large scale data storage as well as database query repositories. Providing computational resources has been guided by the user base as well. As the base grew it became clear that the notebook technology developed by Project Jupyter was highly effective and actually resulted in more people volunteering for the closed beta test. The use of that technology helped bring users into the system.

2.3 Technology Exploration

Many different technologies were tested in 2018. Many options using Amazon Web Services were explored and almost to a point those services were excellent. Local UVA resources were also used and in particular the UVA HPC portal developed out of the VP-IT's office is phenomenal. Collaboration is also underway with the UVA library regarding the discovery component of the Open Data Lab and the implementation of Harvard's Datavese, known locally at UVA as Libra. For version control and sharing purposed GitHub was evaluated.

2.3.1 Amazon Web Services

Cloud computing provides agility that local computational resources do not. To that end we established a contract with Amazon Web Services (AWS) through the third-party vendor DLT solutions. This contract is collectively negotiated and takes advantage of the Internet-2 network. Here is a link to the contract details.

In selecting a cloud resource provider our choice was informed by the needs of the MSDS students at the DSI. The most requested cloud service was AWS and the plurality of job postings that are interested in cloud skills prefer AWS. During 2018 the initial scope of the AWS exploration was for functionality on colocating data and computation. However there was substantial mission creep and now Pete has functionally become the sysadmin for the DSI AWS account.

In the following sections we will breakdown the different services used by the ODL.

Summary of Services Evaluated in 2018

Service	Function	Notes
S3	Object Storage	$30/TB/yr
EC2	Compute Instances	pricing
SageMaker	Project Jupyter	popular interface
IAM	Identity and Access Management	users and groups
API Gateway	Credentialed REST	allows trigger of lambda
Lambda	Serverless Functions	perform miscellaneous tasks
CloudWatch	Log/Monitor/Alarm	more important once scaled

S3

This service provides the cheapest usable storage solution at scale. Furthermore recent policy developments at Amazon now require that other services interact with S3 (object storage) with comparable performace to block storage. That policy is very good for the ODL purposes. It means that cheaper Object Storage can be used for larger and larger datasets without yielding performance of execution. Additionally this means only one storage solution needs to be implemented. We do not need to provision additional block storage for execution operations requiring data migration and costs of duplicate storage.

The S3 systems divides the data into buckets. For this test we treated a bucket as the unit holding data for a particular research project. As a result we have 22 buckets provisioned to accomodate all of our efforts. All buckets on AWS are localized to a region but must have a globally unique name. To that end for phase 1 we have adopted the following convention. Each bucket id begins with 'odl-' and is followed by the four character project id (eg: odl-hmtt).

Summary of S3 Buckets Provisioned for ODL 2018

ID	Project	Notes
odl-bamc	Bourne/Mura Capstone	
odl-bamc-scratch	Bourne/Mura Capstone	scratch space
odl-cbwc		
odl-dome	Dominion Capstone	defunct
odl-hmtt	Healthy Markets	13 TB
odl-hmtt-scratch	Healthy Markets	scratch space
odl-nept	Numismatic Linked Open Data	in development
odl-orci	Educational Open Datasets	in development
odl-pmis		
odl-podc	DSI Communications	
odl-projhects-test	DSI project configurations	
odl-readonly-test	read only test	
odl-scratch-test	scratch space test	
odl-sp19-sys6016	Class materials	
odl-spark-education	Spark Educational Materials	
odl-spark19spds6003-001	Class Materials	
odl-watr	Women and Terrorism Recruitment	
odl-watr-scratch	Women and Terrorism Recruitment	
odl-wiki	Wiki Capstone	
2017-2018-capstone-plos	17/18 capstone	remnant from phase 0
uva-bucket	initial test bucket	remnant from phase 0
846033058400-dlt-utilization	DLT bucket	part of DLT contract

For the example of 'odl-hmtt' this bucket serves the Healthy Markets project. We bill that project PTAO for the service the ODL provides. The mechanism is to pay the DLT contract off the ODL PTAO and then do a cost transfer annually from the Healthy Markets PTAO to the ODL PTAO.

13 TB Data Transfer

When we acquired the Healthy Markets dataset we transferred it from their aws bucket to ours. However this transfer was not trivial. AWS was unwilling to change the bucket ownership from their account to our so we had to copy the data over.

The recommendation was to use the AWS CLI to perform the copy. It functions very similarly to scp from standtard unix systems. However we discovered several interesting features.

- Some of the files were copied to our bucket using an IAM account on their AWS account. Then Healthy Markets deleted that IAM user and as a result we lost control of the files in our bucket. Efforts to regain control are ongoing through DLT solutions.
- Direct bucket to bucket transfer is managed via a serverless operation and does not prototype the transfer. As a result the resources allocated automatically were not sufficient to transfer 13 TB during a human lifetime. We then setup a dedicated server via EC2 and were able to configure the system to complete the transfer at a rate of about 4 TB per day. We did incur cost for operating that server but is was not prohibitive.
- The AWS CLI is not optimized for mass transfer and to copy the data in a reasonable timescale we had to write some scripts to chunk the operation. That work is located on the ODL github page.

EC2

This service allows for provisioning of compute resources. We established lambda functions to automate the create of EC2 instances for projects given a json configurationf file. Here is the link to the code. And here is a sample JSON file:

```
{
   "projectId": "open-data-lab",
   "github": "https://github.com/UVA-DSI/Open-Data-Lab.git",
   "data-bucket": "odl-hmtt",
   "scratch-bucket": "odl-scratch-test",
   "ImageId": "ami-b70554c8",
```

```
   "InstanceType": "t2.nano",
   "email": "lpa2a@virginia.edu",
   "maxNumInstances": "1"
}
```

These EC2 units are monitored by the CloudWatch system and we can configure them to turn on and off as necessary.

The majority of our users use EC2 as their compute engine however they access the compute via auto-provisioning from the SageMaker service.

SageMaker (Project Jupyter)

SageMaker has been the breakout star of the 2018 ODL development phase. Our MSDS students prefer Project Jupyter as their mechanism for using computational and storage resources. To that end Pete went to the JupyterCon in New York City in the fall of 2018. Details of his experience are found here. At this conference Pete was able to speak directly with the creators, developers, and users of Project Jupyter as well as the AWS developers of SageMaker. This service through the power of Project Jupyter is able to democratize the access and management of computational resources. The systems lifts a large cognative load off the user and empowers them to accomplish their goals efficiently without requiring arcane knowledge of computers. Our working metaphor is as follows:

> **The user is able to drive the car and get where they want to go without needing to first learn how to build a transmission.**

This power is the true killer feature of Project Jupyter and is the reason the system is so popular and widely used. Often people we say the ability to break code into cells and use markdown and inline plotting is the killer feature but it is really the lifting of the cognitive load.

IAM

This services governs Identity and Access Management. It breaks down into users and groups. Users are placed into appropriate logical groups and then access policies are assigned to groups. When testing permission settings a dummy IAM user is established and given the same groups as the user

being tested. In that way the sysadmin can see what the user sees and debug/test/prepare the system to the user.

API Gateway

We configured an API gateway to provide a mechanism for the users to trigger lambda functions. The users require credentialing via the standard aws authentication protocol when the post to the Gateway. Details of the system are on the GitHub here.

```
requests.post('https://pish6mpnr0.execute-api.us-east-1.amazonaws.com/alpha-2/vm_s
                        auth=auth,
                        params={'projectID':sys.argv[1]})
```

lambda

This services allows for serverless execution of code. Currently we use it for provisioning of EC2 instances and automation of EC2 management.

CloudWatch

This service is how we monitor the system and record logs.

Architecture Diagrams

Support Plan

For the first year of the ODL we elected to keep the business support plan. In the future we can have a ten percent savings by eliminating this support.

`https://aws.amazon.com/premiumsupport/compare-plans/`

2.3.2 Local UVA - Rivanna and Ivy

The local computational resources at UVA are facilitated through the office of Vice President for IT. That group is dedicated and hard working and provides great resources to the local UVA community. We have established a working relationship with them and discuss technical problems and solutions. Independently we arrived at the utility of Project Jupyter. These solutions are for UVA personnel and their collaborators and as such will not scale

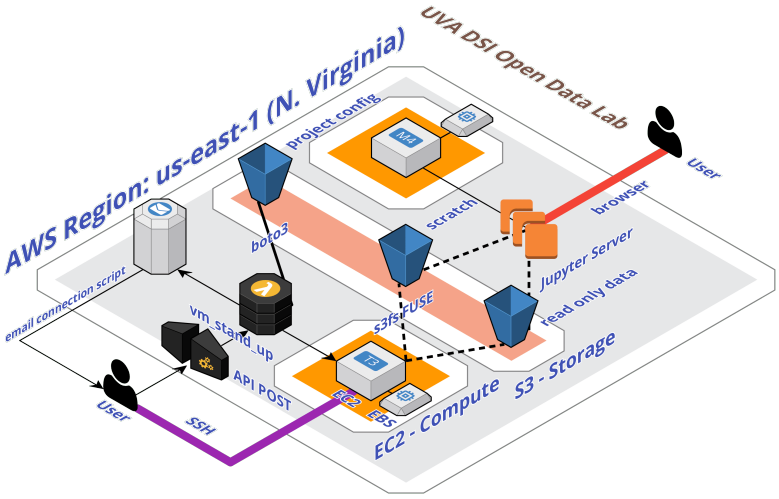


Figure 2.1: Schematic of AWS service configuration

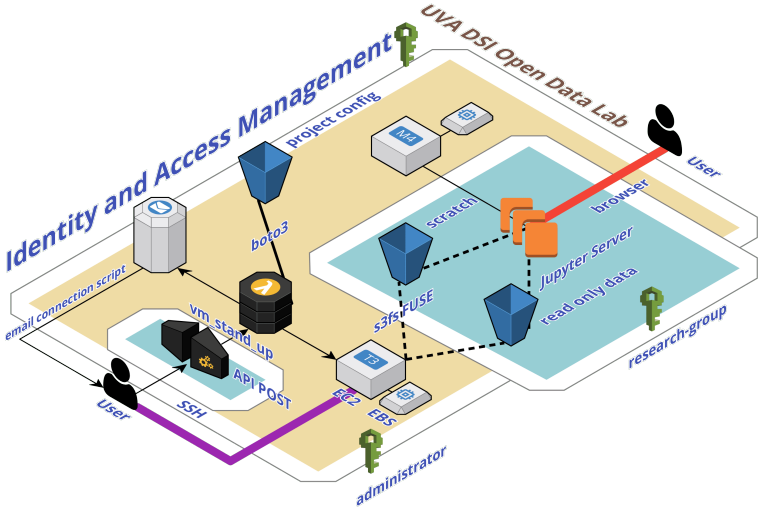


Figure 2.2: AWS IAM configuration

to later phases of the Open Data Lab project. However for the closed and open beta it is a great resource. Furthermore their technical expertise will be invaluable to the Open Data Lab regardless of phase.

2.3.3 GitHub

GitHub is the most broadly adopted cloud platform for version control. Therefore we evaluated it first. The utility for managing repositories is fully mature. The collaborative features focused around the fork and pull request paradigm are excellent. GitHub also has project level capability with issue tracking and team/permission functionality for managing permissions and progress. We have been extremely pleased with the capabilities of GitHub. The only motivation to try other solutions is for the sake of due diligence.

Concerning the acquisition by Microsoft: Many have raised the issue that GitHub may not be the appropriate solution now that Microsoft has acquired GitHub. However the recent track record of Microsoft is to not meddle with projects like GitHub but rather to protect them. Additionally most users use other Microsoft products. What's more the Open Data Lab also relies heavily on Amazon.

2.4 Upcoming Technical Exploration

The following sections describe exploratory work that is on the schedule. There is more to be done beyond this list but not scheduled.

2.4.1 Dataverse

A framework has been outlined to use Dataverse as the discovery mechanism for the Open Data Lab. In this system a metadata entry will be made in the Dataverse containing all of the usual materials. However the final piece with the datafiles will contain pointers to the data and projects within the Open Data Lab. Dataverse is not configured for colocating computation resources with the data resources. The pilot of this test will be with the Libra project from the UVA Library. Currently that system is undergoing an upgrade and once there is a stable release exploration will commence.

2.4.2 Spark

The first scale data solution the Open Data Lab will explore is Spark. Preliminary work so far as been the development of a introductory workshop on the technology (available on the Open Data Lab github repository). A second pedagogical series will be presented early in 2019 and will lead to testing different technical solutions.

2.4.3 SPARQL Endpoint

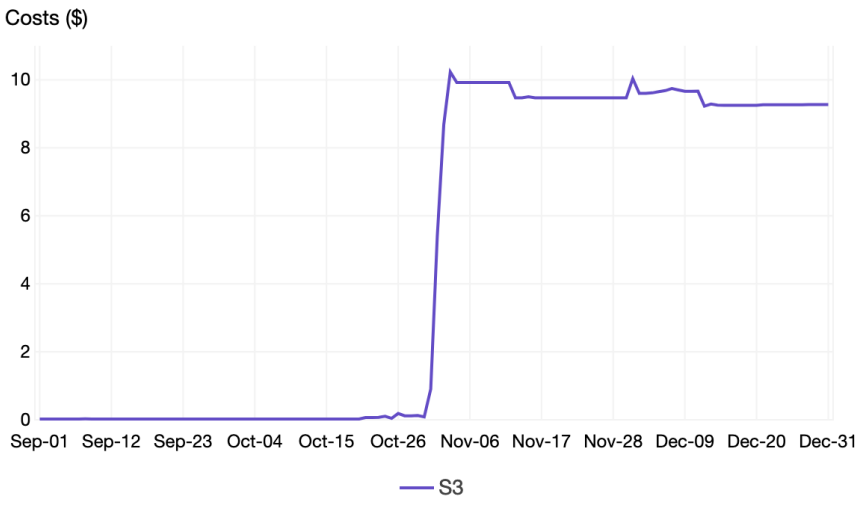
The numismatic dataset will be accessible through a SPARQL endpoint. This exploration is in the early stage and has not matured to the point of evaluation. The next annual report will have a full breakdown of the best way to treat this form of data and delivery.

Chapter 3

Data sets

The Open Data Lab is in the business of hosting data sets with various levels of openness. We encourage all users to make their data as open as practicable. Currently there is a purchased data set that goes by the codename 'Healthy Markets' in the ODL (it contains financial information). And we are bringing a Numismatic data set online. There are also various data sets for student research projects hosted on the open data lab.

At the end of 2018 the total amount of data in the lab was 13.6 TB. This is predominantly from the Healthy Markets Dataset. The following graphic shows the data usage overtime.



3.1 Healthy Markets

The Healthy Markets dataset was purchased by the University under liscense for use by all members of the university. The Open Data Lab is responsible for storage and providing access to the dataset. Narjessadat Seyeditabari (Narges Tabari) is responsible for facilitating research on the dataset and is the primary user of the dataset.

The data set is 13 TB and is stored in an S3 bucket named 'odl-hmtt'. This bucket is private to users of the Open Data Lab and is accessed through a SageMaker instance. To copy the data from it's source at Healthy Markets we used the AWS CLI and it took several days with a total cost of $1 for REST actions and $17 for the EC2 instance to manage the transfer. The current cost to store the data set is $290 per month. These costs are covered by the Healthy Markets PTAO.

Section 4.3 contains the report on research activities with this data set.

3.2 Numismatic

Dr. Ethan Gruber from the American Numismatic Society is in possession of a Linked Data Set and wants to make it open with the Open Data Lab. He is interested in establishing a SPARQL endpoint for the dataset. An S3 bucket (odl-nept) has been provisioned to store the data for the project and we will be making the data set open in 2019.

Chapter 4

Research

4.1 Bourne/Mura Capstone

4.2 DSI Wiki

4.3 Healthy Markets

Chapter 5

Education

A key component of openness is making resources usable. This idea falls in line with the idea that openness and accessibility are part of the same mission. As a result the educational component of the Open Data Lab is vital to the suces of the project. There are two main thrusts in this endeavor. The first, which was piloted in 2018, is the production of educational materials and methods. The second part is the development of communication paradigms. This year two workshops were produced and delivered as part of the closed beta test. The first focused on the scale data protocol spark and the second the version control tool GitHub.

5.1 Spark Workshop

This workshop was designed as an introduction to spark. The goals were:

- Teach how to get started
- Build comfort
- Teach how to get answers to further questions

The topics covered included linking to a spark context, reading in data via dataframes, manipulating the data, and making a fundamental calculation. To power the workshop the attendees were given credentials on a Amazon SageMaker notebook. One of the features of this approach is that the whole workshop has access to the same environment. Everyone sees the same implementation of the software and hardware. There is no cumbersome overhead

in getting set up. The requirements are a web browser and access to the internet. Furthermore the single notebook environment leads to a very useful pedagocial capability. When a student encounters and error in their code the instructor can load their notebook on the main display in the room. In real time and in full view of everyone the instructor can debug and teach the whole class. This is a vast improvement over the current popular method of hovering over a single learners station. It enables every person in the room to see what is going on and maintain their level of engagement. This experience was very positive for the learners and the feedback to this approach was superlative. Previous versions of spark training was done with Databricks resources and there were several drawbacks precipitating the switch to Amazon.

- The environment is not shared between the workshop participants and the instructor
- Every learner independently established their own cluster and there is substantial lag
- Materials must be imported in Databricks format (.dbc) instead of more universal jupyter notebook format (.ipynb).

Resources:

- Databricks based workshop can be found at: https://github.com/alonzi/spark
- Amazon Sakemaker based workshop can be found at: https://github.com/alonzi/spark-intro
- Next generation materials will be incorporated into the Open Data Lab repository at: https://github.com/UVA-DSI/Open-Data-Lab/tree/master/education

The cost to operate and instructional environment is $0.0464 per hour. For this workshop we ran the environment for one week at a cost of approximately $10.

5.2 GitHub Workshop

The Open Data Lab was invited to present GitHub to the Archaeology Department of the Thomas Jefferson Foundation (aka Monticello). We developed a workshop to explain the fundamentals of version control and present

a work-flow for beginning users. The different user archetypes were also discussed. One of the major burdens to version control use is that it comes from the computer superuser community. Most of the software is developed using a terminal based interlace (CLI). However in today's research world many one computer superusers interact with code and other materials that benefit from a version control workflow. The major benefit from GitHub is the browser based interface. This implementation shifts substantial pieces of cognative load off the user. This shift enables the user to focus on developing their work rather than on the bookkeeping of version controling their work. At the same time it makes it easy for the developers to take advantage of the version control benefits. There is substantial room to further develop materials for different user archeytypes. This workshop focused on a research group. We will strive to identify other archetypes and develop materials to suit those needs. This workshop was taught straight from the GitHub repository itself. That was a natural fit given the subject matter. But it also demonstrated several very useful pieces of GitHub as a teaching medium, which will be discussed in 5.3.

Resources: https://github.com/UVA-DSI/Open-Data-Lab/tree/master/education/GitHub

5.3 Using GitHub as a Teaching Medium

Both of the workshops taught under the Open Data Lab project used GitHub as the repository for materials. This has several benefits.

- GitHub provides a URL and free hosting for resources
- Subsequent changes to the materials are stored under version control thus allowing the actual materials presented to be recovered
- Any learner who wants to suggest improvements to materials can implement a pull request

The decision to put the materials in GitHub was one of necessity since the Open Data Lab GitHub page serves as the repository for all Open Data Lab resources. GitHub by default provides a URL for every item stored in the repository and presents the README file of a repo. This presentation is automatically rendered from markdown. Wikipedia has demonstrated the success of using markdown for content presentation but to enumerate some

key features here. The document is organized, hyperlinkable, figures are easily embeded, and it seamlessly renders text alongside code and mathematical formulae.

5.4 Plans for 2019

The Open Data Lab is scheduled to teach several workshops in 2019:

- A five part series on Spark for the DSI MSDS cohort
- A three part series for the UVA Library on Python and Machine Learning
- Likely another three part series for the UVA Library in Fall 2019

potential collab with COS?

Chapter 6

Vital Metrics

6.1 AWS usage

6.2 FTE analysis

6.3 Budget

6.3.1 Funding

funded by the Data Science Institute healthy markets project funded off-budget for ODL